

# Natural Neighbor-based Clustering Algorithm with Density Peaks

Dongdong Cheng, Qingsheng Zhu, Jinlong Huang, Lijun Yang  
Chongqing Key Laboratory of Software Theory & Technology,  
College of Computer Science, Chongqing University, Chongqing, China  
Email: qszhu@cqu.edu.cn

**Abstract**—Clustering analysis has been widely used in many areas such as astronomy, bioinformatics, and pattern recognition. In 2014, Rodriguez proposed an algorithm based on the idea that cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher density. But the density relies on cutoff distance, which might be affected by large statistical error, and the algorithm does not suit the clustering problem of multi-scale data. In this paper, a new neighbor concept Natural Neighbor is proposed. Natural neighbor-based density, is simple and well reflects the data distribution without any parameters. Then, we extend each cluster from its center by searching natural neighbors of points in this cluster, and we define extension rules to determine the cluster boundary. The experiment results show our algorithm is more effective on multi-scale data.

## I. INTRODUCTION

Clustering is an important data analysis method with unsupervised learning process. The objective of clustering is to classify elements into categories on the basis of their similarity. So, it can be applied in many fields such as machine learning, biology and recognition.

Some different clustering strategies have been proposed in Ref. 1, but clustering remains a difficult problem because of the inherent vagueness in the definition of a cluster and the difficulty in defining an appropriate similarity measure and objective function. Partition-based method is one of primary clustering algorithms that use an iterative control strategy to optimize an objective function, such as K-means[2] and K-medoids[3], but the result is affected by the initial center selection. Hierarchical clustering algorithms recursively find nested clusters either in agglomerative (bottom-up) mode or in divisive (top-down) mode. Chameleon is a representative hierarchical method of integrating two approaches[4]. Density-based methods such as DBSCAN[5] are able to detect nonspherical clusters, but choosing appropriate parameters for these algorithms can be nontrivial. In order to process large dataset more efficiently, the grid-based clustering algorithms[6] have been proposed and there are some improve algorithms[7], [8]. Spectral clustering[9] has become one popular clustering algorithm, which makes use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. There also exists some graph clustering with spectral method[10].

In 2007, a new clustering algorithm by passing messages between data points (AP)[16] was proposed. AP takes as

input measures of similarity between pairs of data points. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges. However, AP clustering algorithm cannot directly specify the final cluster number. K-AP[17] and DAAP[18] are the improved algorithm of AP algorithm.

In order to discover nonspherical clusters, some new algorithms are proposed such as nearest neighbor-based clustering algorithms[11], [12], [13]. On Science in 2014, a clustering algorithm [14] by finding density peaks adopts an alternative way to automatically find the correct number of clusters and discover nonspherical clusters. The algorithm first finds cluster centers with the decision graph, then each remaining point is assigned to the same cluster as its nearest neighbor of higher density, and last finds the point of highest density within its border region for each cluster, the points of the cluster whose density is higher than  $\rho_b$  are considered part of the cluster core, and the others are considered part of the cluster halo. However, it still needs to set the value cutoff distance, and it does not solve the clustering problem of multi-scale data.

In this paper, we introduce the concept of Natural Neighbor, which adaptively obtains the number of neighbors without any parameters, and propose Natural Neighbor-based clustering algorithm with density peaks (NaNDP). The rest of this paper is organized as follows: the second part briefly reviews cluster\_dp algorithm; the third part introduces the idea of our method NaNDP; the forth part verifies the efficiency of NaNDP algorithm using artificial data set, real-world data set and Olivetti Face Database, respectively; and finally, we make a conclusion on our work and point out the direction for further research.

## II. CLUSTERING BY FAST SEARCH AND FIND OF DENSITY PEAKS

Clustering by fast search and find of density peaks in Ref. 14 is an efficient clustering algorithm, here we call it cluster\_dp, which is based on the assumption that cluster centers are surrounded by neighbor with lower local density and that they are at a relatively large distance from any points with a higher local density. For each data point  $i$ , the local density  $\rho_i$  is defined as

$$\rho_i = \sum_j \chi(d_{ij} - d_c)$$

where  $\chi(x) = 1$  if  $x < 0$  and  $\chi(x) = 0$  otherwise, and  $d_c$  is a cutoff distance. In fact,  $\rho_i$  is equal to the number of points that are closer than  $d_c$  to point  $i$ .

The distance  $\delta_i$  is measured by computing the minimum distance between the point  $i$  and any other point with higher density:

$$\delta_i = \min_{j: \rho_j > \rho_i} (d_{ij})$$

For the point with highest density  $\delta_i = \max_j (d_{ij})$

According to the definition,  $\delta_i$  is much larger than typical nearest neighbor distance only for points that are local or global maxima in the density. Cluster centers are points with the value of  $\delta_i$  is anomalously large.

Decision graph shows the plot of  $\delta_i$  as a function of  $\rho_i$  for each point. The points with relatively high  $\rho$  and  $\delta$  are selected to be cluster centers, and points with relatively high  $\rho$  and low  $\delta$  are considered as outliers. Fig. 1 has shown a dataset and its decision graph. From Fig. 1(b), two points with relative high density  $\rho$  and distance  $\delta$  are away from the rest of points. According to the previous describe, the two points are cluster centers and can be selected easily.

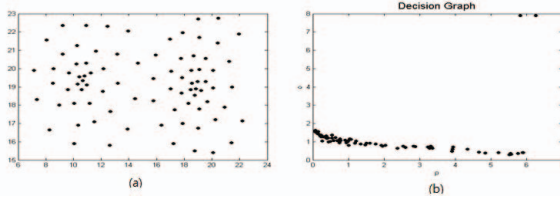


Fig. 1. A dataset and its decision graph

However, when the density of two clusters is significantly different, it will be hard to set an appropriate cutoff distance, which will influence clustering results, just as shown in Fig. 2. Besides, the clustering process of cluster\_dp that assigns each remaining point to the same cluster as its nearest neighbor of higher density, makes it hard to find clusters with manifold data structure, which is shown in Fig. 3, there are two line clusters, but cluster\_dp cannot correctly recognize the two clusters. The black star points are cluster centers found by decision graph and points with the same color belong to the same cluster which is detected by cluster\_dp. The two points enclosed by red circle are points with local maximum density, but their  $\delta$  value is smaller than the two black star points, and their nearest neighbor with higher density are respectively the points pointed by arrows, which leads to the wrong clustering result. In this paper, we present Natural Neighbor-based clustering algorithm with density peaks, which computes density without any parameters and extends each cluster from its center by searching nearest neighbors of points in this cluster.

### III. NATURAL NEIGHBOR-BASED CLUSTERING ALGORITHM

In this paper, we introduce Natural Neighbor. A new concept of density based on Natural Neighbor is proposed. On

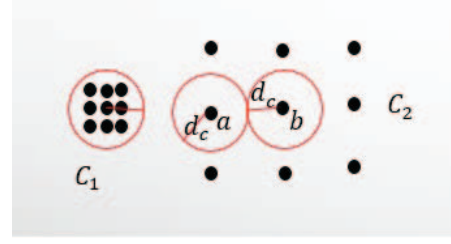


Fig. 2. Illustration of great density variations, the value of cutoff distance is hard to set.

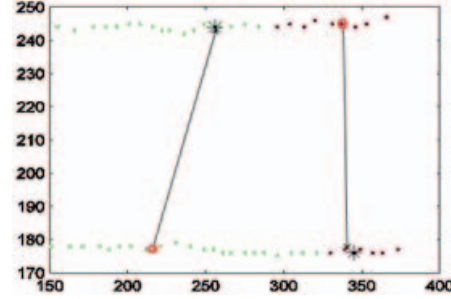


Fig. 3. The data set with manifold structure that cluster\_dp cannot process, the black star points are cluster centers found by decision graph, and the points enclosed by red circle are with local maximum density and assigned to the cluster the nearest neighbor with higher density (pointed by the arrows) belong to, resulting in wrong clustering.

the basis of the new concept of density and cluster\_dp, we present Natural Neighbor-based Clustering Algorithm with Density Peaks (NaNDP). First, we use the new defined density construct decision graph, and the points with high density and high distance in the decision graph are selected as cluster centers. After cluster centers have been found, clusters are expanded from centers on the Maximum Neighbor Graph (MNG) according to some rules. If point doesn't satisfy the rules, it will not be clustered in the second step. And last, the rest points without being clustered are classified according to their neighbors which have been clustered.

#### A. Natural Neighbor

Natural Neighbor (NaN) is a new concept of neighbor. The concept originates from the reality that the number of one's real friends should be the number of how many people are taken him or her as friends and he or she take them as friends at the same time. When all persons (including outliers) in a community have natural neighbor, the community will be harmony. For data objects, object  $y$  is one of the Natural Neighbor of object  $x$  if object  $x$  considers  $y$  to be a neighbor and  $y$  considers  $x$  to be a neighbor at the same time. In particular, the object lying in sparse region should possess small number of neighbors, whereas object lying in dense region should possess large number of neighbors. The key idea of Natural Neighbor is that the object lying in sparse region should possess little energy, whereas object lying in dense region should possess higher power. The whole computational

procedure of Natural Neighbor can be automatically fulfilled without any parameters. If the formation of K-Nearset Neighbor is regarded as an active neighbor searching procedure, then the forming of Natural Neighbor is completely passive.

The search cost of KNN and RNN for each object in the database is huge. So we introduce the KD-tree [21] into the Natural Neighbor searching. The Natural Neighbor searching algorithm is described in Algorithm 1.

**Definition 1** (Natural Neighbor): Based on the Natural Neighbor searching algorithm, the following will be shown, if point  $x$  belongs to the neighbor of point  $y$  and  $y$  belongs to the neighbor of point  $x$ , then  $x$  is called as  $y$ 's Natural Neighbor (NaN), and  $y$  is Natural Neighbor of  $x$ .

**Definition 2** (Natural Eigenvalue  $sup_k$ ): NaN implies that each point has different number of neighbors, for any point  $i$ , its number of neighbors is  $nb(i)$ . But NaN has an average number of neighbors, called  $sup_k$  which is natural characteristic value. The formula of computing  $sup_k$  is following.

$$sup_k = \{r | \forall x \exists y (y \neq x \cap x \in NN_r(y)) \text{ or } \forall x$$

$$(|RNN_r(x)| = 0) = (|RNN_{r-1}(x)| = 0)\}$$

Here,  $x$  and  $y$  are the points of dataset.  $NN_r(y)$  is the  $r$ -th neighborhood of  $y$ .  $RNN_r(y)$  is the  $r$ -th reverse nearest neighbor of  $y$ .

**Definition 3** (Saturated Neighborhood Graph SNG): The graph constructed by linking  $sup_k$  nearest neighbors of each point is called as Saturated Neighborhood Graph (SNG).

**Definition 4** (Maximum Neighborhood Graph MNG): The graph constructed by linking  $\max\{nb(i), sup_k\}$  nearest neighbors of each point is called as Maximum Neighborhood Graph (MNG). The connectivity of MNG is better than SNG.

---

• **Algorithm1: NaN-Searching(dataset)**

---

- 1) Initializing:  $r=1, nb(i)=0, NN_0(i) = \phi, RNN_0(i) = \phi$
  - 2)  $kdtree=creatKDTree(dataset)$  //creat a KD-tree
  - 3) Use  $kdtree$  to find the  $r$ -th neighbor  $y$  for each data point  $x$ .
    - a.
      - a)  $Rnb(y) = Rnb(y)+1$
      - b)  $NN_r(x) = NN_{r-1}(x) \cup \{y\}$
      - c)  $RNN_r(y) = RNN_{r-1}(y) \cup \{x\}$
  - 4) Compute the number of data point  $x$  that  $nb(x)=0$  a.
    - a) If the number don't changed for 2 times  
goto step5
    - b) else  
 $r = r + 1$  and goto step3
  - 5)  $sup_k = r$
  - 6) output the  $sup_k, nb, NN(i)$
- 

In Algorithm 1,  $nb(y)$  represents the times that point  $y$  is contained by the neighborhood of other points, which is the number of  $y$ 's reverse neighbor.  $NN_r(x)$  is the  $r$  nearest-neighbors of  $x$ .  $RNN_r(y)$  is the  $r$ -reverse neighbors of  $y$ ;  $sup_k$  is Natural Eigenvalue. Since KD-tree is introduced into NaN-Searching, the time complexity of NaN-searching algorithm is  $O(N \cdot \log N)$ .  $N$  is the number of data in dataset.

## B. Density based on Natural Neighbor

A concept of density based on  $k$  nearest neighbor is proposed in Ref. 13, which helps describe the dissimilarity between points, and we use the similar definition to compute density. The key is to determine the appropriate  $k$  value. According to the concept of Natural Neighbor, different points have different number of neighbors, indicating their contribution of density degree or neighborhood structures to neighbors. In order to compute density more accurately, we select the maximum number of natural neighbors that is  $\max\{nb\}$  as the  $k$  value.

**Definition 5** (The density of a given point  $i$ ): Let  $Max\_nb = \max\{nb\}$ , the density of point  $i$  is computed as follow:

$$\rho_i = \frac{Max\_nb}{\sum_{j \in N(i, Max\_nb)} dist(i, j)}$$

Where  $N(i, k)$  is the  $k$  nearest neighbors of point  $i$ , and  $dist(i, j)$  is the distance between point  $i$  and  $j$ .

## C. Extending cluster from center on MNG

The extending process is on the assumption that the boundary between clusters is sparse, and density of the point on the boundary is lower than others, if its neighbors' density is greater, the neighbors possibly belong to other clusters. In order to determine the cluster boundary, we define extension rule. We assume Point  $P$  is the extending point, and Point  $Q$  is one of the neighbors of  $P$ , which has not been extended. If  $Q$  meets one of the following rules,  $Q$  will be extended.

- (1) The density of  $Q$  is lower than  $P$ .
- (2) The density of  $Q$  is larger than  $P$ , and the nearest neighbor of  $Q$  belongs to the cluster which  $P$  belongs to.

Fig. 4 has shown the extending process,  $D$  and  $H$  are cluster centers determined by decision graph. First, we extend cluster from  $D$ , the neighbors of  $D$  are  $A, C$ .  $A, C$  satisfy Rule (1), then  $A, C$  are extended by  $D$ . Similarly,  $B$  is extended by  $A$  and  $E, G$  are extended by  $H$ . Thus  $A, B, C$  and  $D$  are classified in the first cluster. Then we extend from  $E$ , the density of  $F$  is larger than  $E$ , according to Rule (2), if  $F$ 's nearest neighbor is  $G$ , and  $G$  is already extended by  $H$ , then  $F$  will be extended by  $E$ . However, if  $F$ 's nearest neighbor is  $I$ , and  $I$  has not been extended,  $F$  will not be extended by  $E$ . Thus  $F, I$  will be assigned to none of the clusters. After extending from center on MNG, although some points are assigned to none of the clusters, we can guarantee that the clustered points are correctly classified. The rest points are clustered according to their neighbors that have been clustered. The NaN algorithm is detailed as Algorithm 2.

---

• **Algorithm2: NaNP(dataset)**

---

- 1) Initializing:  $NCLUST=0, cl(i)=0$  //  $NCLUST$  is the cluster number and  $cl(i)$  is the cluster label of point  $i$ .
- 2) Using the NaN-Searching algorithm to construct the MNG, and obtain the  $K=\max(nb)$
- 3)  $\rho = \text{ComputeDensity}(K)$
- 4)  $\delta = \text{ComputeDistance}(\rho)$

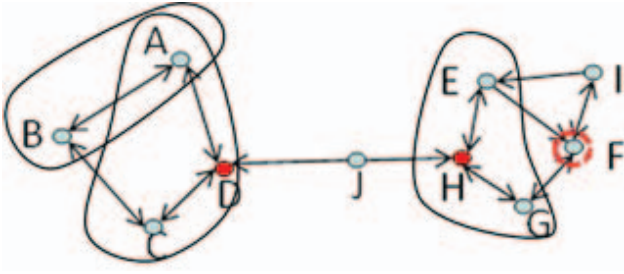


Fig. 4. The extending process on MNG. The red points are cluster centers determined by decision graph. the descending order of density is :  $H > G > F > E > I > D > C > J > A > B$ . Points surrounded by curves are extended according to Rule (1). Point F will be extended by Rule (2). If its nearest neighbor is G, F will be extended. If its nearest neighbor is I, F will not be extended.

- 5) ConstructDecisionGraph(rho,delta)
- 6) Centers=searchCenters()
- 7) For each center C in Centers (b)
  - a) Q=queue()
  - b) Q.add(C)
  - c) NCLUST=NCLUST+1
  - d) While Q is not empty (ii)
    - i) x=Q.remove()
    - ii) cl(x)=NCLUST
    - iii) For each neighbor y in x on MNG,if y satisfy the Rules then Q.add(y)
  - e) minrho(C)=min{rho(x)|x ∈ C}
- 8) Clustering for the rest points according to their neighbors.
- 9) Return cl

The proposed method calculates the density without any parameters according to the results of NaN-Searching algorithm, which is one of its strengths. And it extends cluster from center on MNG making it more adaptive to the manifold data structure. The main cost of time is that obtaining the neighbors of each points in Algorithm 1. Hence, the time complexity of NaNDP is  $O(N \cdot \log N)$ .

#### IV. EXPERIMENTS AND PERFORMANCE EVALUATION

##### A. Assessment of clustering performance

We evaluate the clustering performance using two criteria. The first one, clustering error (CE)[15]measures the error rate. Usually, the serial number of every cluster will be resorted, after the clustering results are obtained, e.g. the first cluster of original data set may be specified as second cluster by algorithm. Thus, we need to construct a displacement mapping function and match the serial number obtained by clustering and real cluster label, one by one. CE index is based on this mapping relation, and the calculation formula is as follows:

$$CE = 1 - \frac{1}{n} \sum_{i=1}^n \delta(y_i, map(c_i))$$

Where  $y_i$  is real cluster label,  $c_i$  is the serial number obtained by clustering, and  $\delta(x, y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$  is discriminate function. CE indicates the situation that has the least wrong classification among all possible mapping.  $CE \in [0, 1]$ , the smaller the value of CE, means the better the clustering performance of the algorithm.

The second performance evaluation criterion is Normalized Mutual Information (NMI). The NMI is defined as

$$NMI(X, Y) = \frac{MI(X, Y)}{\sqrt{H(X)H(Y)}}$$

Where  $MI(X, Y)$  is the mutual information between two random variables  $X$  and  $Y$ , and  $H(*)$  is the random variable entropy, which is used for normalizing the mutual information to be in the range of  $[0, 1]$ . Given ground-truth label vector and clustering result, the  $NMI$  measures how good the clustering results is, with respect to ground-truth.  $NMI = 1$  means the clustering result is perfect and  $NMI = 0$  means the clustering result is useless. Other value between 1 and 0 measures the quality of the clustering result.

##### B. Clustering on artificial data sets

In order to test the performance of our method NaNDP and cluster\_dp algorithm, we choose seven challenging artificial data sets illustrated in Fig. 5. The first 4 datasets are used in Ref. 14, and we use them to prove we also can do well in their datasets. The other three are datasets that cannot be processed by cluster\_dp. Here, when clustering with cluster\_dp, we don't consider finding the cluster halo.

In experiment 1, we use four datasets shown in Fig.5(a) (b) (c) and (d). In the first two datasets, there are 1000 and 4000 points generated from a probability distribution respectively. Fig. 6 presents the result of the dataset in Fig. 5(a), and the first row is the decision graph and clustering results of cluster\_dp, and the second is of our method. It's the same with Fig. 7 which depicts the result of the dataset in Fig. 5(b). Table 1 shows the run time of cluster\_dp and our method. From the clustering results, we can see if we don't consider noise points, our algorithm has the similar effect and speed on the datasets experimented by cluster\_dp. In the other two datasets, the first

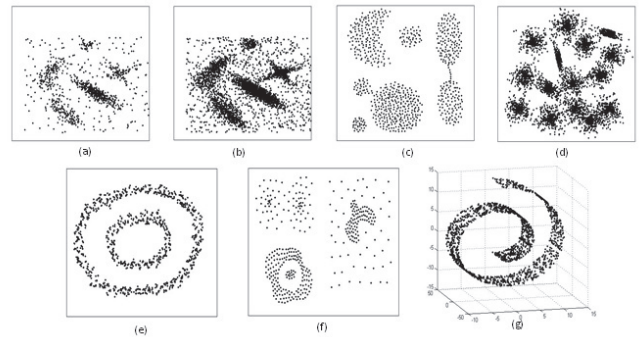


Fig. 5. Original artificial data sets.



TABLE I  
THE RUN TIME OF CLUSTER\_DP AND OUR METHOD (S).

Method	Dataset in Fig. 2(a)	Dataset in Fig. 2(b)
Cluster_dp	281.0	280.81
NaNDP	876.0	875.74

one contains 7 clusters with concave and convex shapes and the second one comprises 15 spherical clusters. Fig. 8 has shown the clustering results of datasets in Fig. 5 (c) and (d). The first row is the results of cluster\_dp, and the second is our method. From the results we can see that both of the two methods obtain good clustering results, which proves that our method is as good as cluster\_dp for some datasets and our method can do well on datasets the cluster\_dp dose.

In experiment 2, we compare our method with cluster\_dp on the other three datasets shown in Fig. 5(e) (f) and (g). The dataset in Fig. 5(e) consists of two circle clusters which the small one is in the big one, and has 1022 data points. There are 399 points and six clusters in the other dataset, especially, in which there is a dense cluster embedded in a sparse cluster which is regarded as noise points here. Fig. 5(g) is the three dimension dataset with 1600 points, which is generated by Swiss Roll model and composed of two Swiss roll Clusters and each one has 800 points. Fig. 9 unveils clustering results of cluster\_dp in first row and our method in second row. The dark points of the second last picture are recognized as noise points. Seen from the clustering results, our method correctly finds the clusters in the three datasets, however, cluster\_dp is not suitable for the dataset that contains manifold clusters. The reason is that each remaining point is assigned to the same cluster as its nearest neighbor of higher density in cluster\_dp. Our method clustering with nearest neighbors does better work than cluster\_dp especially for datasets with manifold structure like that in Fig. 5(e) (f) and (g).

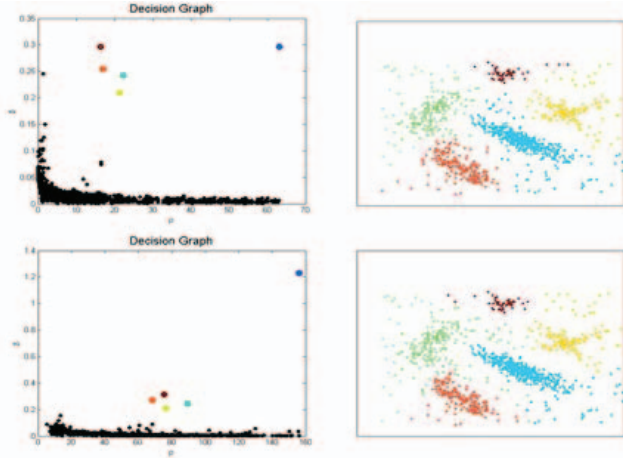


Fig. 6. The decision graphs and clustering results of dataset in Fig 2(a). The first row is the result of cluster\_dp and the second row is the result of our method.

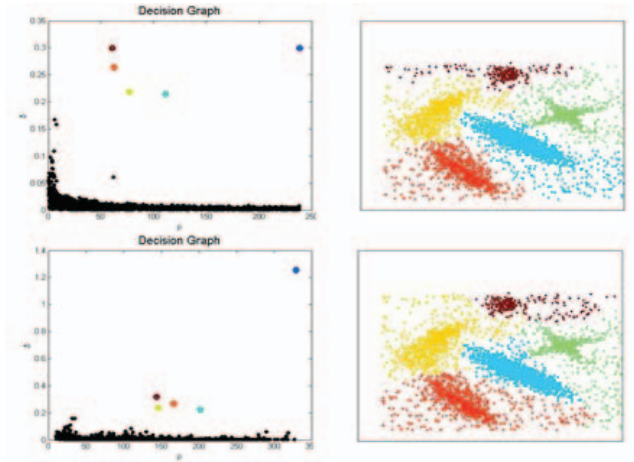


Fig. 7. The decision graphs and clustering results of dataset in Fig 2(b). The first row is the result of cluster\_dp and the second row is the result of our method.

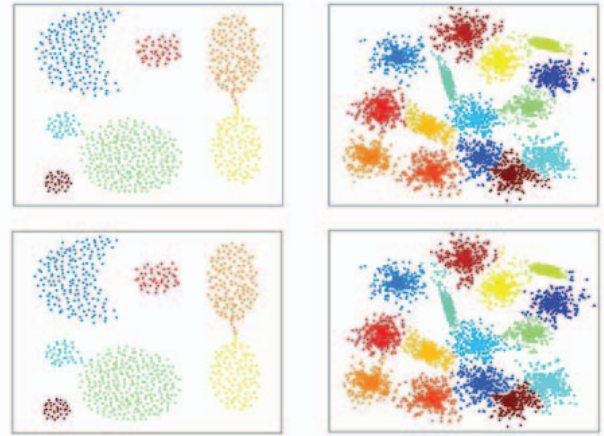


Fig. 8. Clustering results of datasets in Fig 2(c) and 2(d). 1st row are the results of cluster\_dp and 2nd row are the results of our method.

### C. Clustering on real data sets

To further demonstrate the effectiveness of our algorithm, we compare our algorithm to AP [16] and cluster\_dp [14] on several benchmarking real-world data sets from UCI. The characteristics of these data sets are shown in Table 2. In the experiment, we exploit CE index and NMI index as measurement criterion. The results of our experimental are illustrated in Table 3. The configuration of the computer used in our experiment is as follows: processor is Intel Core i5 2.80GHZ; memory size is 4 GB; programming environment is MATLAB R2013a.

According to Table 3, CE index and NMI index are almost consistent. For AP algorithm, the accuracy of clustering is not high, and usually it produces bad clustering results, since it is hard to control the number of clustering. Cluster\_dp

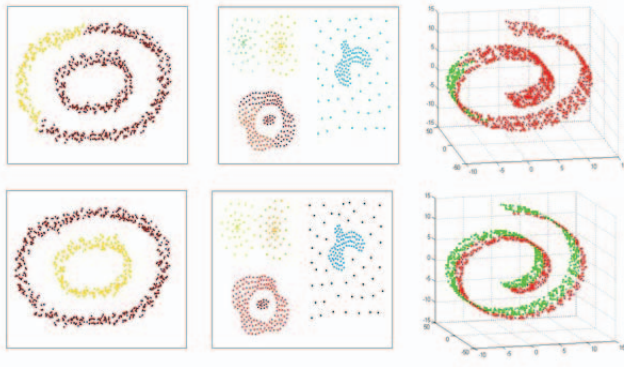


Fig. 9. Clustering results of cluster\_dp and our method on the other three data sets. 1st row are the results of cluster\_dp and 2nd row are the results of our method.

TABLE II  
DATA CHARACTERISTICS OF REAL DATA SETS

Data set	samples number	attributes number	clusters number
Wine	178	13	3
Iris	150	4	3
Control	600	60	6
Breast	699	10	2

and NaNDP can make good use of decision graph and the prior knowledge and detect the correct number of clusters successfully. However, as for cluster\_dp, the density might be affected by large statistical errors, and the clustering process doesn't suit for multi-scale data, so the performance is poor. The clustering results of NaNDP are significantly better than those of cluster\_dp algorithm, indicating that natural neighbor-based density efficiently describe the distribution characteristic of the data and the clustering process is more effective when processing manifold data. The bold data in each line is the best result. As for the four real data sets, it is obvious that our algorithm is even better than cluster\_dp algorithm.

#### D. Clustering on Olivetti Face Database

We also applied our algorithm to the Olivetti Face Database [17]. The Olivetti Face Database contains 400 face images from 40 persons, taken at different times and varying the

lighting, facial expressions and facial details. The size of each image is  $92 \times 112$  pixels. We use the first 100 images, that is, 10 clusters of the database to do the experiment. The similarity between two images, denoted as  $S(A,B)$ , is computed by following equation.

$$S(A, B) = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{(\sum_m \sum_n (A_{mn} - \bar{A})^2)(\sum_m \sum_n (B_{mn} - \bar{B})^2)}}$$

Here  $A$  and  $B$  are the subjects of Olivetti Face Database.  $A_{mn}$  and  $B_{mn}$  represent the pixels of the two images. The value of  $S$  is scaled to  $[0,1]$ . The bigger the value of  $S$  is, the more similar the two images are. The distance between two images, denoted as  $d(A,B)$ , is computed as following:

$$d(A, B) = 1 - S(A, B)$$

For such a small set, the search of natural neighbor is unavoidably affected by noise objects, making the value of  $sup_k$  larger than it should be. When extending cluster from center, we reduce the value of  $sup_k$ . Here, we also sparse the data points by the plot of  $\gamma_i = \rho_i \delta_i$  sorted in decreasing order, and then select a minimum  $\gamma_i$  for choosing cluster centers. Fig. 10 has shown the clustering results when considering the noises. The density of noises is much smaller than that of the normal. When extending each cluster from the center, we will denote the minimum density as the cluster density threshold. The points with lower density will be considered as noises, and will be assigned to none of the clusters. Fig. 11 is clustering results of our method when we do not consider the noises, and every points are assigned to one of the clusters. The results show that NaNDP correctly selects 10 centers marked with red squares. When the objects with lower density are denoted as noises, all of the 10 clusters are pure, and we can accurately detect more images than cluster\_dp[14]. When all of the images are clustered, there are only three images enclosed with red box in Fig. 11 are assigned to the wrong cluster.

Through above experiments and analysis, it is obvious that introducing Natural Neighbor-based density avoids large statistical errors caused by inappropriate cutoff distance and NaNDP does as good as or better work than cluster\_dp, especially for data set with manifold structure.

TABLE III  
CLUSTERING RESULTS ON REAL DATA SETS

Data set		AP	cluster_dp	NaNDP
Wine	CE	0.6629	0.0393	<b>0.0337</b>
	NMI	0.6008	0.8571	<b>0.8782</b>
	Cluster number	14	3	3
Iris	CE	0.6067	0.1667	<b>0.0400</b>
	NMI	0.6632	0.7403	<b>0.96</b>
	Cluster number	11	3	3
Control	CE	0.7017	<b>0.3700</b>	0.3733
	NMI	0.5809	0.7205	<b>0.7219</b>
	Cluster number	24	6	6
Breast	CE	0.8984	0.1517	0.1330
	NMI	0.3893	0.4245	<b>0.4580</b>
	Cluster number	11	2	2

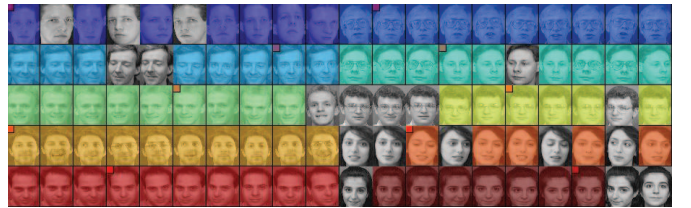


Fig. 10. Cluster results of NaNDP on the Olivetti Face Database when we consider noises. Faces with the same color belong to the same cluster, whereas gray images are noises detected by our method and not assigned to any cluster. Cluster centers are labeled with red squares.



Fig. 11. Clustering results of NaNDP on the Olivetti Face Database when we don't consider noises, and every image is assigned to one of the clusters. Faces with the same color belong to the same cluster. Cluster centers are labeled with red squares. The faces enclosed by red box are assigned to the wrong cluster.

## V. CONCLUSIONS AND FUTURE SCOPE

Cluster\_dp is a relative efficient algorithm, which computes density according to a cutoff distance set by users, and can fast and correctly find nonspherical clusters of some datasets by discovering cluster centers according to density peaks. But the density computed by cutoff distance is unavoidably affected by large statistic errors and the process of clustering makes it does not fit for datasets containing ring structure clusters or other manifold datasets. In this paper, we propose a method which computes density based on Natural Neighbor. Natural neighbor automatically adapts to the distribution of datasets and gets the number of points neighbors. We use Max\_nb nearest neighbors of points to compute density which is as effective as cluster\_dp when drawing decision graph. After cluster centers have been found, the remaining points are clustered based on MNG, making our method has more advantages in terms of discovering clusters with manifold structure.

However, our method has disadvantages in discovering noises points. It cannot exactly find noises in some datasets, and most noises are assigned to a certain cluster. This is what we do in future to find a good way to deal with noises while clustering, and avoiding the influence of noises.

## ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (61272194) and National Natural Science Foundation of China (61502060).

## REFERENCES

- [1] Jain A K. Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 2010, 31(8): 651-666.
- [2] MacQueen J. Some methods for classification and analysis of multivariate observations. *Proc. of 5th Berkeley symposium on mathematical statistics and probability*. 1967, 1(14): 281-297.
- [3] Kaufman L, Rousseeuw P J. *Finding groups in data: an introduction to cluster analysis* (John Wiley & Sons, 2009).
- [4] Karypis G, Han Euihong, Kumar V. Chameleon: Hierarchical Clustering Using Dynamic Modeling. *Computer*. 1999, 32(8): 68-75.
- [5] Ester M, Kriegel H P, Sander J. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*. 1996, 96(34): 226-231.
- [6] Schikuta E. *Grid-Clustering: A hierarchical clustering method for very large data sets*. In Technical Report TR-CRPC No.93358, Center for Research on Parallel Computation. Rice University, 1993.
- [7] Esfandani G, Saayadi M and Namadchian A. GDCLU: a new grid density based CLUstering algorithm. *Proc. of 13th ACIS Conf*. pp102-107, 2012.

- [8] Dou W, Hu J. A half-split grid clustering algorithm by simulating cell division. *2014 IEEE International Joint Conference on Neural Networks*. pp2183-2189, 2014.
- [9] Von Luxburg U. A tutorial on spectral clustering. *Statistics and computing*. 2007, 17(4): 395-416.
- [10] Chen W Y, Song Y, Bai H, et al. Parallel spectral clustering in distributed systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. 2011, 33(3): 568-586.
- [11] Chen X. A new clustering algorithm based on near neighbor influence. *Expert Systems with Applications*, 2015.
- [12] Ritter GX, Nieves-Vzquez JA, Urcid G. A simple statistics-based nearest neighbor cluster detection algorithm. *Pattern Recognition*, 2015, 48(3): 918-932.
- [13] Wan JX, Zhu QS, Lei DJ. Outlier detection based on transitive closure[J]. *Intelligent Data Analysis*, 2015, 19(1): 145-160.
- [14] Rodriguez A, Laio A. Clustering by fast search and find of density peaks. *Science*, 2014, 344(6191): 1492-1496.
- [15] Jordan F, Bach F. Learning spectral clustering. *Adv. Neural Inf. Process. Syst*, 2004, 16: 305-312.
- [16] Frey B J, Dueck D. Clustering by passing messages between data points[J]. *science*, 2007, 315(5814): 972-976.
- [17] Samaria F S, Harter A C. Parameterisation of a stochastic model for human face identification. *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*. IEEE, 1994: 138-142.
- [18] Zhang X, Wang W, Norvag K, et al. K-AP: generating specified K clusters by efficient affinity propagation. *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010: 1187-1192.
- [19] Jia H, Ding S, Meng L, et al. A density-adaptive affinity propagation clustering algorithm based on spectral dimension reduction. *Neural Computing and Applications*, 2014, 25(7-8): 1557-1567.
- [20] Bentley J L. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 1975, 18(9): 509-517.